

Oracle Inequalities for High-dimensional Prediction

JOHANNES LEDERER^{1,2}, LU YU¹, and IRINA GAYNANOVA³

¹*Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
E-mail: ledererj@uw.edu*

²*Department of Statistics
University of Washington
Box 354322
Seattle, WA 98195
E-mail: luyu92@uw.edu*

³*Department of Statistics
Texas A&M University
458D Blocker
College Station, TX 77843
E-mail: irinag@stat.tamu.edu*

Abstract The abundance of high-dimensional data in the modern sciences has generated tremendous interest in penalized estimators such as the lasso, scaled lasso, square-root lasso, elastic net, and many others. However, the common theoretical bounds for the predictive performance of these estimators hinge on strong, in practice unverifiable assumptions on the design. In this paper, we introduce a new set of oracle inequalities for prediction in high-dimensional linear regression. These bounds hold irrespective of the design matrix. Moreover, since the proofs rely only on convexity and continuity arguments, the bounds apply to a wide range of penalized estimators. Overall, the bounds demonstrate that generic estimators can provide consistent prediction with any design matrix. From a practical point of view, the bounds can help to identify the potential of specific estimators, and they can help to get a sense of the prediction accuracy in a given application.

Keywords: Oracle inequalities, high-dimensional regression.

1. Introduction

Oracle inequalities are the standard theoretical framework for measuring the accuracy of high-dimensional estimators [8]. Two main benefits of oracle inequalities are that they hold for finite sample sizes and that they adapt to the underlying model parameters. Oracle inequalities are thus used, for example, to compare estimators and to obtain an idea of the sample size needed in a specific application.

For high-dimensional prediction, there are two types of oracle inequalities: so-called fast rate bounds and so-called slow rate bounds. Fast rate bounds hold for near orthogonal designs and bound the prediction error in terms of the sparsity of the regression vector. Such bounds have been derived for a number of methods, including the lasso, the square-root lasso, and their extensions to grouped variables, see [3, 8–10, 15, 32] and others. Slow rate bounds, on the other hand, hold for any design and bound the prediction error in terms of the penalty value of the regression vector. Such bounds have been developed for lasso estimators [18, 20, 22, 23], but have not received much attention for other methods. However, unlike the unfortunate naming suggests, slow rate bounds are of great interest. In particular, slow rate bounds are not inferior to fast rate bounds, quite in contrast [13, 16]: (i) Slow rate bounds hold for any design, while fast rate bounds impose strong and in practice unverifiable assumptions on the correlations in the design. (ii) Even if the assumptions hold, fast rate bounds can contain unfavorable factors, while the factors in slow rate bounds are small, global constants. (iii) Also in terms of rates, slow rate bounds can outmatch even the most favorable fast rate bounds. See [13] and references therein for a detailed comparison of the two types of bounds. To avoid confusion in the following, we will use the terms *penalty bounds* instead of slow rate bounds and *sparsity bounds* instead of fast rate bounds.

In this paper, we develop new penalty bounds for prediction in high-dimensional linear regression. These oracle inequalities hold for any sample size, design, and noise distribution. Moreover, our proof technique applies to a very general family of objective functions, including square-root lasso and group square-root lasso, while the standard proof technique for penalty bounds is limited to lasso-type estimators. As a bonus, our approach improves the standard bounds for lasso-type estimators by a factor 2.

The organization of the paper is as follows. Below, we introduce the setting and notation. In Section 2, we state the general result. In Section 3, we specialize this result to specific estimators. In Section 4, we conclude with a brief discussion. The proofs are deferred to the Appendix.

The ordering of the following sections is geared towards readers that wish to dive into the technical aspects right away. For getting a first overview instead, one can have a quick glance at the model and the estimators in Displays (1) and (2), respectively, and then skip directly to the examples in Section 3.

Setting and Notation

We consider linear regression models of the form

$$Y = X\beta^* + \varepsilon \tag{1}$$

with outcome $Y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n \times p}$, regression vector $\beta^* \in \mathbb{R}^p$, and noise vector $\varepsilon \in \mathbb{R}^n$. Our goal is prediction, that is, estimation of $X\beta^*$. We allow for general design matrices X , that is, we do not impose conditions on the correlations in X . Moreover, we allow for general noise ε , that is, we do not restrict ourselves to specific distributions for ε .

We are particularly interested in high-dimensional settings, where the number of parameters p rivals or even exceeds the number of samples n . As needed in such settings, we assume that the regression vector β^* has some additional structure. This structure can be exploited by penalized estimators, which are the most standard methods for prediction in this context. We thus consider estimators of the form

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{g(\|Y - X\beta\|_2^2) + \text{pen}(\lambda, \beta)\} ,$$

where g is some real-valued link function, the mapping $\text{pen}(\lambda, \beta) : \mathbb{R}^k \times \mathbb{R}^p \rightarrow \mathbb{R}$ accounts for the structural assumptions, and λ is a vector-valued tuning parameter. More specifically, we assume that the assumptions on β^* can be captured by (semi-)norms. The corresponding estimators then read

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\} . \quad (2)$$

To derive results that are specific enough to be useful in concrete examples, we impose some additional conditions: The function $g : \mathbb{R} \rightarrow [0, \infty)$ satisfies $g(0) = 0$, is continuous and strictly increasing on $[0, \infty)$, and is continuously differentiable on $(0, \infty)$ with strictly positive and non-increasing derivative $g'(x) := \frac{d}{dy}g(y)|_{y=x}$. Moreover, the function $\mathbb{R}^n \rightarrow [0, \infty) : \alpha \mapsto g(\|\alpha\|_2^2)$ is assumed to be strictly convex. The most common examples for the link function are $g(x) = x$ and $g(x) = \sqrt{x}$.

Next, we assume that the entire penalty is in the form of a composite norm. First, we assume that the tuning parameter $\lambda := (\lambda_1, \dots, \lambda_k)^\top$ is in $(0, \infty)^k$. We then assume that the matrices $M_1, \dots, M_k \in \mathbb{R}^{p \times p}$ satisfy $\bigcap_{j=1}^k \text{Ker}(M_j) = \{\mathbf{0}_{p \times p}\}$, where Ker denotes the kernel of a matrix. In the simplest cases, the matrices M_j equal the identity matrix. In general, however, these matrices allow for the incorporation of complex structural assumptions. For example, group structures can be modeled by taking each M_j equal to a block-diagonal matrix with some of the blocks equal to zero.

The norms $\|\cdot\|_{q_j}$ with $q_j \geq 1$ are finally the regular ℓ_{q_j} -norms on \mathbb{R}^p . Their dual norms are denoted by $\|\cdot\|_{q_j}^*$, and it holds that $\|\cdot\|_{q_j}^* = \|\cdot\|_{p_j}$ for $p_j \in [1, \infty]$ such that $1/p_j + 1/q_j = 1$. Since $\|\cdot\|_{q_j}$ is a norm, and since the rows of the matrices M_j span the entire \mathbb{R}^p , the penalty is indeed a norm. (One could generalize the penalty further; for example, one could replace the ℓ_{q_j} -norms by general semi-norms. However, we omit this for the sake of clarity.)

We finally need to ensure that single variables are not over-penalized if they are subject to more than one penalty term. For this, we first denote by A^+ the Moore-Penrose pseudoinverse of a matrix A . We then note that by the rank assumption on the matrices M_1, \dots, M_k , there are projection matrices $P_1, \dots, P_k \in \mathbb{R}^{p \times p}$ such that

$$\sum_{j=1}^k P_j M_j^+ M_j = \mathbf{I}_{p \times p} . \quad (3)$$

The projection matrices enter the “empirical process” terms in the definition of the oracle tuning parameters and in the prediction bounds. Our results hold for any P_1, \dots, P_k that satisfy the above equality; however, appropriate choices are needed to obtain sharp bounds. In generic examples, the choice of P_1, \dots, P_k is straightforward: if $k = 1$ (see, for example, the lasso, square-root lasso, and fused lasso) or if the row spaces of the matrices M_1, \dots, M_k are disjoint (see, for example, the group lasso with non-overlapping groups), one can select $P_1, \dots, P_k = I_{p \times p}$. More generally, if $k > 1$ and some variables are penalized twice (see, for example, the group lasso with overlapping groups), slightly more complicated choices lead to optimal bounds.

Finally, we exclude non-generic distributions. More specifically, we assume that $Y \neq \mathbf{0}_n$ and $\min_{j \in \{1, \dots, k\}} \|(XP_j M_j^+)^\top \varepsilon\|_{q_j}^* > 0$ with probability one. This implies in particular that $g'(\|Y - X\hat{\beta}^\lambda\|_2^2) > 0$ with probability one, see Lemma A.3 in the Appendix.

2. General Result

We now state the general oracle inequality. For this, we recall that oracle inequalities are optimized by oracle tuning parameters. The standard oracle tuning parameters for the lasso, for example, are of the form $a\|X^\top \varepsilon\|_\infty$, where the factor $a \geq 2$ depends on the specific type of oracle inequality ($a = 2$ for the standard penalty bounds; $a > 2$ for standard sparsity prediction or estimation bounds, see [4, 12] and others). To obtain a general definition of oracle tuning parameters for our bounds, we derive the following result.

Lemma 2.1 (Existence). *With probability one, there is a tuning parameter $\lambda \in (0, \infty)^k$ such that*

$$\frac{\lambda}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} = (\|(XP_1 M_1^+)^\top \varepsilon\|_{q_1}^*, \dots, \|(XP_k M_k^+)^\top \varepsilon\|_{q_k}^*)^\top.$$

This proof of existence ensures that oracle tuning parameters can be properly defined for arbitrary estimators of the form (2). If $g : x \mapsto x$, Lemma 2.1 can be verified easily. In particular, the above equation simplifies to $\lambda = 2\|X^\top \varepsilon\|_\infty$ for the lasso. In general, however, the statement is more intricate, and there might be several tuning parameters that satisfy the equality. The proof is then based on continuity arguments and Brouwer’s fixed-point theorem, see Appendix.

We can now define oracle tuning parameters as follows.

Definition 2.1 (Oracle tuning parameters). *We define the oracle tuning parameter $\bar{\lambda} \in (0, \infty)^k$ as any λ that satisfies the equality in Lemma 2.1. For brevity of notation, we then also set $\bar{\beta} := \hat{\beta}^{\bar{\lambda}}$, which is the estimator at this oracle tuning parameter.*

Lemma 2.1 ensures that an oracle tuning parameter exists. We show in the next section, where we look at the lasso, square-root lasso, and other examples, that Definition 2.1 specializes correctly.

We are finally ready to state the main result.

Theorem 2.1 (Penalty bound). *With probability one, the estimator (2) with oracle tuning parameter according to Definition 2.1 satisfies the prediction bound*

$$\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{2}{n} \sum_{j=1}^k \|(XP_j M_j^+)^\top \varepsilon\|_{q_j}^* \|M_j \beta^*\|_{q_j}.$$

This oracle inequality provides bounds for the prediction errors of the estimators (2). Our proofs are based on convexity and continuity, making the results considerably more general and slightly sharper than the standard results, see the following section.

We briefly highlight five main features of Theorem 2.1 (much of this becomes more lucid in the context of specific examples, see the next section): First, the bounds involve the oracle tuning parameters to the power one (cf. Definition 2.1 and Corollary 2.1 below) and hold for any design matrix X . Moreover, the bounds contain the penalty value of the regression vector. Hence, the bounds are penalty bounds. Second, the oracle inequality holds for any distribution of the noise ε . Third, the bounds hold for any sample size n ; in particular, the bounds are non-asymptotic. Fourth, the bounds become smaller if the correlations in X become larger, cf. [16] (in contrast, sparsity bounds are smallest for orthogonal design). Fifth, the link function g appears in the oracle tuning parameter but not in the prediction bound. This last, interesting point clarifies the role of the link function: its purpose is basically to reshuffle the tuning parameter path. One can relate this observation to the lasso/square-root lasso discussion, see the examples section and [3, 21].

By construction, the oracle tuning parameters lead to the sharpest bounds. However, corresponding bounds can be derived also for larger tuning parameters. For illustration, we consider $k = 1$, $g : x \mapsto x$, and take as oracle tuning parameter $\bar{\lambda}$ the largest λ that satisfies the conditions of Lemma 2.1 (we show in the proof of Lemma 2.1 that with probability one, there is a bounded set that contains all candidate λ). We then write $\lambda = \lambda_1$, $M = M_1$ and derive the following oracle inequality.

Corollary 2.1 (General tuning parameters). *If $k = 1$ and $g : x \mapsto x$, then with probability one, all estimators $\hat{\beta}^\lambda$ with $\lambda \geq \bar{\lambda}$ satisfy the prediction bound*

$$\frac{1}{n} \|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 \leq \frac{\lambda}{n} \|M \beta^*\|_{q_j}.$$

We omit the proof, since it follows the same lines as the proof of Theorem 2.1. For general link functions g , one can use $\lambda/2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \approx \lambda/2g'(\|\varepsilon\|_2^2)$ to obtain similar results. For $k > 1$, one has to introduce an ordering on \mathbb{R}^k ; we omit details to avoid digression.

Before heading to the examples in the next section, we conclude this part with additional pointers to related literature. First, we refer to [13, 16] for in-depth comparisons of penalty bounds and sparsity bounds for the lasso, and for studies on the influence of the design matrix on the oracle tuning parameter. Second, we refer to [11, 12] for

approaches to how oracle tuning parameters can be mimicked in practice, and we refer to [3, 14, 21, 27] and the square-root lasso examples below for approaches to make the oracle tuning parameters independent of unknown model aspects.

3. Examples

We now state the results of Theorem 2.1 for specific estimators of the form (2). Some of these, such as the (group) square-root lasso, are not amenable to the standard proof technique, and have thus not been equipped with penalty bounds. It is straightforward to derive the results from Theorem 2.1, but the calculations are clunky at points. For ease of presentation, we thus leave the detailed calculations to the reader.

Lasso The lasso [28] is defined as

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \}.$$

Since $k = 1$ and $M_1 = P_1 = I_{p \times p}$ in this example, the corresponding oracle tuning parameter is $\bar{\lambda} = 2\|X^\top \varepsilon\|_\infty$, and the bound reads

$$\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{2}{n} \|X^\top \varepsilon\|_\infty \|\beta^*\|_1.$$

For example, if $\varepsilon_1, \dots, \varepsilon_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and $(X^\top X)_{jj} = n$ for all $j \in \{1, \dots, p\}$, it holds that $\bar{\lambda} \approx \sigma \sqrt{n \log(p)}$ and $\|X(\beta^* - \bar{\beta})\|_2^2/n \lesssim \sigma \sqrt{\log(p)/n} \|\beta^*\|_1$.¹ Importantly, the rate $1/\sqrt{n}$ is the optimal rate - unless further assumptions on the design matrix X are imposed [13, Example 1 and Proposition 2].

Comparing with the standard penalty bounds for the lasso, see [20, Eq. (2.3) in Theorem 1] and [16, Equation (3)], we find that the oracle tuning parameters coincide, but that our bounds (see also Corollary 2.1) improve on the standard bounds by a factor 2. To see this, recall that the standard proof technique uses that the lasso objective function is minimal at $\hat{\beta}^\lambda$, so that in particular

$$\|Y - X\hat{\beta}^\lambda\|_2^2 + \lambda \|\hat{\beta}^\lambda\|_1 \leq \|Y - X\beta^*\|_2^2 + \lambda \|\beta^*\|_1. \quad (4)$$

Invoking the model (1), this yields

$$\|X\beta^* - X\hat{\beta}^\lambda\|_2^2 \leq 2\langle X\hat{\beta}^\lambda - X\beta^*, \varepsilon \rangle + \lambda \|\beta^*\|_1 - \lambda \|\hat{\beta}^\lambda\|_1.$$

Hölder's inequality and the triangle inequality then lead to

$$\|X\beta^* - X\hat{\beta}^\lambda\|_2^2 \leq 2\|X^\top \varepsilon\|_\infty (\|\hat{\beta}^\lambda\|_1 + \|\beta^*\|_1) + \lambda \|\beta^*\|_1 - \lambda \|\hat{\beta}^\lambda\|_1.$$

¹The wiggles indicate that we are interested only in the rough shapes and neglect constants, for example.

Hence, for $\lambda = \bar{\lambda} = 2\|X^\top \varepsilon\|_\infty$, we find

$$\|X\beta^* - X\hat{\beta}^\lambda\|_2^2 \leq 4\|X^\top \varepsilon\|_\infty \|\beta^*\|_1.$$

In our proofs, we replace the “zeroth order” starting point (4) with a “first order” starting point, see Lemma A.4. This approach leads to more general results and a smaller factor.

Square-root/scaled lasso The square-root lasso [3] reads

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{\|Y - X\beta\|_2 + \lambda \|\beta\|_1\}.$$

In our framework, $k = 1$ and $M_1 = P_1 = I_{p \times p}$. Consequently, the corresponding oracle tuning parameter satisfies $\bar{\lambda} = \|X^\top \varepsilon\|_\infty / \|Y - X\hat{\beta}^\lambda\|_2$, and the prediction bound reads

$$\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{2}{n} \|X^\top \varepsilon\|_\infty \|\beta^*\|_1.$$

Hence, the prediction bounds are the same as for the lasso, but the oracle tuning parameters differ. The crux of the square-root lasso, and similarly, the scaled lasso [27], is that their oracle tuning parameters can be essentially independent of the noise variance. For example, if $\varepsilon_1, \dots, \varepsilon_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and $(X^\top X)_{jj} = n$ for all $j \in \{1, \dots, p\}$, it holds that $\bar{\lambda} \approx \|X^\top \varepsilon\|_\infty / \|\varepsilon\|_2 \approx \sqrt{\log(p)}$, which is independent of σ . Since σ is typically unknown in practice, the square-root/scaled lasso can thus facilitate the tuning of λ .

Slope estimator The slope estimator [6] can be written as

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^p \omega_j |\beta|_{(j)} \right\},$$

where $|\beta|_{(j)}$ denotes the j th largest entry of β in absolute value, $\omega_1 \geq \dots \geq \omega_p > 0$ is a non-increasing sequence of weights, and $\lambda > 0$ is a tuning parameter. A promising case for the slope estimator is the one where $\varepsilon_1, \dots, \varepsilon_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and $(X^\top X)_{jj} = n$ for all $j \in \{1, \dots, p\}$. The weights can then be chosen as $\omega_j := 2\sigma\sqrt{n \log(2p/j)}$ in the spirit of the Benjamini-Hochberg procedure [6, 26], and a theoretically justified choice of the tuning parameter is $\lambda > 4 + \sqrt{2}$ [2, Equation (2.5)]. In particular, this choice works in the sense of Corollary 2.1, and one finds the bound $\|X(\beta^* - \bar{\beta})\|_2^2/n \lesssim \lambda \sigma \sqrt{\log(p)/n} \|\beta^*\|_1$, which coincides with the bounds above. Similar considerations apply to the oscar penalty [7].

Elastic net The elastic net [35] reads

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{\|Y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2\}.$$

This is not directly in the form (2). However, one can use the usual trick writing the estimator as a lasso with augmented data, cf. [17, Lemma 1]. Using $M_1 = P_1 = I_{p \times p}$,

our results then hold for any tuning parameters that satisfy $\lambda_1 = 2\|X^\top \varepsilon - \lambda_2 \beta^*\|_\infty$. For example, we can set $\bar{\lambda}_2 = \arg \min_{\lambda_2} \|X^\top \varepsilon - \lambda_2 \beta^*\|_\infty$ and $\bar{\lambda}_1 = 2\|X^\top \varepsilon - \bar{\lambda}_2 \beta^*\|_\infty$. The corresponding bound then reads

$$\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{2}{n} \|X^\top \varepsilon - \bar{\lambda}_2 \beta^*\|_\infty \|\beta^*\|_1 \leq \frac{2}{n} \|X^\top \varepsilon\|_\infty \|\beta^*\|_1.$$

Similar results hold if, for example, the settings with the normal noise vectors described in the two examples above apply and $\bar{\lambda}_2 = \mathcal{O}(\sqrt{n})$. The main intent of the elastic net is to improve variable selection. However, our results show that the elastic net with well-chosen tuning parameters also has similar penalty guarantees for prediction as the lasso.

Lasso and square-root lasso with group structures The estimators considered so far are based on a simple notion of sparsity. In practice, however, it can be reasonable to assume more complex sparsity structures in the regression vector β^* . Estimators that take such structures into account include the group lasso [34], group square-root lasso [9], hierarchical group lasso [5], and sparse group lasso [25]. They all fit our framework.

In contrast to the examples above, the matrices $M_1, \dots, M_k, P_1, \dots, P_k$ play a non-trivial role in these examples. On a high level, the matrix M_j specifies which variables are incorporated in the j th group. If groups overlap, the matrix P_j specifies which parts of $X^\top \varepsilon$ are attributed to the oracle tuning parameter $\bar{\lambda}_j$. For example, if the m th variable is in the j th and l th group, the matrices P_j and P_l can be chosen such that the corresponding element $(X^\top \varepsilon)_m$ is part of either $\bar{\lambda}_j$ or $\bar{\lambda}_l$ and not in both of them.

As an illustration, let us consider the group lasso with non-overlapping groups:

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^k \|\beta_{G_j}\|_2 \right\}.$$

Here, G_1, \dots, G_k is a partition of $\{1, \dots, p\}$ and $(\beta_{G_j})_i := \beta_i \mathbb{1}\{i \in G_j\}$. To put the estimator in framework (2), one can either generalize the estimator to incorporate possibly different tuning parameters for each group, set M_j to $(M_j)_{st} = \mathbb{1}\{s = t, s \in G_j\}$, and then choose a dominating tuning parameter, or one can directly extend our results (as mentioned earlier) to arbitrary norm penalties. In any case, the projection matrices/matrix equal the identity matrix. Moreover, the oracle tuning parameter is $\bar{\lambda} = 2 \max_{l \in \{1, \dots, k\}} \|(X^\top \varepsilon)_{G_l}\|_2$, and the bound is

$$\frac{1}{n} \|X(\beta^* - \bar{\beta})\|_2^2 \leq \frac{2}{n} \max_{l \in \{1, \dots, k\}} \|(X^\top \varepsilon)_{G_l}\|_2 \sum_{j=1}^k \|\beta_{G_j}^*\|_2.$$

Trend filtering and total variation/fused penalty Trend filtering [19, 29] reads

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \beta\|_2^2 + \lambda \|M\beta\|_1 \right\},$$

where for given $l \in \{1, 2, \dots\}$, the matrix $M \in \mathbb{R}^{p \times p}$ is defined as $M := \underbrace{D \times \dots \times D}_{l \text{ times}}$ with $D \in \mathbb{R}^{p \times p}$ given by

$$D_{ij} := \begin{cases} -1 & \text{if } i < p \text{ and } i = j \\ 1 & \text{if } i < p \text{ and } i = j - 1 \\ 0 & \text{otherwise} \end{cases}.$$

The oracle tuning parameter is $\bar{\lambda} = 2\|M^{+\top}\varepsilon\|_\infty$, and the bound is²

$$\frac{1}{n}\|\beta^* - \bar{\beta}\|_2^2 \leq \frac{2}{n}\|M^{+\top}\varepsilon\|_\infty\|M\beta^*\|_1.$$

In the case $l = 1$, the estimator becomes

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \|Y - \beta\|_2^2 + \lambda \sum_{j=2}^p |\beta_j - \beta_{j-1}| \right\},$$

which corresponds to the total variation [24] and fused lasso penalizations [30]. Moreover, one can check that the Moore-Penrose inverse of $M = D$ is then given by D^+ with entries

$$D_{ij}^+ := \begin{cases} (j-p)/p & \text{if } i \leq j < p \\ j/p & \text{if } i > j, j < p \\ 0 & \text{if } j = p \end{cases}.$$

The oracle tuning parameter is $\bar{\lambda} = 2\|D^{+\top}\varepsilon\|_\infty$, and the bound is

$$\frac{1}{n}\|\beta^* - \bar{\beta}\|_2^2 \leq \frac{2}{n}\|D^{+\top}\varepsilon\|_\infty\|D\beta^*\|_1.$$

4. Discussion

Sparsity bounds have been derived for many high-dimensional estimators. In this paper, we complement these bounds with corresponding penalty bounds. Which type of bounds is sharper depends on the underlying model. As a general rule, penalty bounds improve with increasing correlations in the design matrix, while sparsity bounds deteriorate with increasing correlations and are eventually infinite once the design matrix is too far from an orthogonal matrix [13].

As a consequence, without making assumptions on the design, and for a wide range of penalized estimators, our results imply non-trivial rates of convergence for prediction. This is of direct practical relevance, since the assumptions inflicted by sparsity bounds

²Unlike assumed earlier, $\text{Ker}(M) \neq \{\mathbf{0}_{p \times p}\}$ in this example, but one can replace matrix M by the invertible matrix $M + \epsilon \mathbf{I}_{p \times p}$ and then take the limit $\epsilon \rightarrow 0$.

are often unrealistic in applications and, in any case, depend on inaccessible model parameters and thus cannot be verified in practice. For example, sparsity bounds for the lasso have been derived under a variety of assumptions on X , including RIP, restricted eigenvalue condition, and compatibility condition, see [31] for an overview of these concepts. Results from random matrix theory show that these assumptions are fulfilled with high probability if the data generating process is “nice” (sub-Gaussian, isotropic, ...) and the sample size n is large enough, see [33] for a recent result. Unfortunately, in practice, the data generating processes often do not seem sufficiently nice, see [1], for example, and the sample sizes can be small. Moreover, even if the conditions are satisfied, fast rate bounds can contain very large factors and are then only interesting from an asymptotic point of view.

Acknowledgments

We sincerely thank Mohamed Hebiri for his insightful comments on a draft of the paper.

References

- [1] BALDING, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7** 781–791.
- [2] BELLEC, P., LECUÉ, G. and TSYBAKOV, A. (2016). Slope meets Lasso: improved oracle bounds and optimality. *arXiv:1605.08651*.
- [3] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.
- [4] BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732.
- [5] BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.* **41** 1111.
- [6] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. (2015). SLOPE - adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103.
- [7] BONDELL, H. and REICH, B. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123.
- [8] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Series in Statistics.
- [9] BUNEA, F., LEDERER, J. and SHE, Y. (2014). The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms. *IEEE Trans. Inform. Theory* **60** 1313–1325.
- [10] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194.

- [11] CHÉTELAT, D., LEDERER, J. and SALMON, J. (2014). Optimal Prediction in High Dimensions. *arXiv:1410.5014*.
- [12] CHICHIGNOUD, M., LEDERER, J. and WAINWRIGHT, M. (2014). Tuning Lasso for sup-norm optimality. *arXiv:1410.0247*.
- [13] DALALYAN, A., HEBIRI, M. and LEDERER, J. (2014). On the Prediction Performance of the Lasso. *To appear in Bernoulli*.
- [14] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statistical Science* **27** 500–518.
- [15] HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall.
- [16] HEBIRI, M. and LEDERER, J. (2013). How Correlations Influence Lasso Prediction. *IEEE Trans. Inform. Theory* **59** 1846–1854.
- [17] HEBIRI, M. and VAN DE GEER, S. (2011). The Smooth-Lasso and other $\ell_1 + \ell_2$ -penalized methods. *Electron. J. Stat.* **5** 1184–1226.
- [18] HUANG, C., CHEANG, G. and BARRON, A. (2008). Risk of Penalized Least Squares, Greedy Selection and L1 Penalization for Flexible Function Libraries. Manuscript.
- [19] KIM, S. J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). ℓ_1 Trend Filtering. *SIAM review* **51** 339–360.
- [20] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329.
- [21] LEDERER, J. and MÜLLER, C. (2014). Don’t fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. *arXiv:1404.0541*.
- [22] MASSART, P. and MEYNET, C. (2011). The Lasso as an ℓ_1 -ball model selection procedure. *Electron. J. Stat.* **5** 669–687.
- [23] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771.
- [24] RUDIN, L., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena* **60** 259–268.
- [25] SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22** 231–245.
- [26] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068.
- [27] SUN, T. and ZHANG, C. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.
- [28] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.
- [29] TIBSHIRANI, R. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323.
- [30] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B* **67** 91–108.
- [31] VAN DE GEER, S. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392.

- [32] VAN DE GEER, S. and LEDERER, J. (2013). The Lasso, correlated design, and improved oracle inequalities. *IMS Collections* **9** 303–316.
- [33] VAN DE GEER, S. and MURO, A. (2014). On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.* **8** 3031–3061.
- [34] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.
- [35] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.

Appendix A: Proofs

We start with four auxiliary results, Lemmas A.1-A.4. We then prove Lemma 2.1 and Theorem 2.1. Figure 1 depicts the dependence structure of the results.

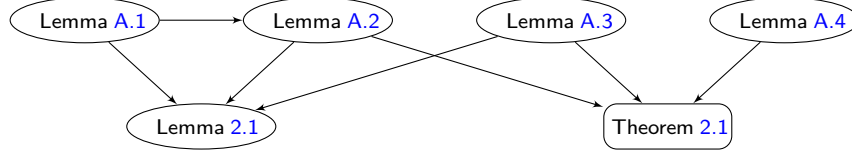


Figure 1. Dependencies between the results. For example, the arrow between Lemma A.1 and Lemma 2.1 depicts that the proof of Lemma 2.1 makes use of Lemma A.1.

A.1. Auxiliary Lemmas

Lemma A.1. For any $\hat{\beta}^\lambda, \tilde{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j}\}$ and $\alpha \in [0, 1]$, it holds that $X\hat{\beta}^\lambda = X\tilde{\beta}^\lambda$ and

$$\alpha \hat{\beta}^\lambda + (1 - \alpha) \tilde{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

Lemma A.2. Let \mathbb{R}^k be equipped with the Euclidean norm, and \mathbb{R} be equipped with the absolute value norm. Then, the function

$$(0, \infty)^k \rightarrow \mathbb{R} \\ \lambda \mapsto g(\|Y - X\hat{\beta}^\lambda\|_2^2)$$

and the function

$$(0, \infty)^k \rightarrow \mathbb{R} \\ \lambda \mapsto \|Y - X\hat{\beta}^\lambda\|_2^2$$

are both continuous.

Lemma A.3. With probability one, it holds that $\|Y - X\hat{\beta}^\lambda\|_2^2 > 0$ and $g'(\|Y - X\hat{\beta}^\lambda\|_2^2) > 0$ for any tuning parameter $\lambda \in (0, \infty)^k$.

Lemma A.4. *Suppose the function*

$$\begin{aligned} \mathbb{R}^p &\rightarrow \mathbb{R} \\ \beta &\mapsto f(\beta) \end{aligned}$$

is convex and $\mathbf{0}_p \in \partial_\beta f(\beta) \Big|_{\beta=\bar{\beta}}$ for a vector $\bar{\beta} \in \mathbb{R}^p$. Then, for every $\kappa \in \partial_\beta f(\beta) \Big|_{\beta=\bar{\beta}}$ and $\tilde{\beta} \in \mathbb{R}^p$, it holds that

$$\kappa^\top (\tilde{\beta} - \bar{\beta}) \geq 0.$$

A.2. Proofs of the Auxilliary Lemmas

Proof of Lemma A.1. The case $\alpha \in \{0, 1\}$ is straightforward, so that we consider a given $\alpha \in (0, 1)$. We first show that $X\hat{\beta}^\lambda = X\tilde{\beta}^\lambda$. Since the function $\alpha \mapsto g(\|\alpha\|_2^2)$ is strictly convex by assumption, it follows for any vectors $a, \tilde{a} \in \mathbb{R}^n$ that

$$g(\|\alpha a + (1 - \alpha)\tilde{a}\|_2^2) \leq \alpha g(\|a\|_2^2) + (1 - \alpha)g(\|\tilde{a}\|_2^2)$$

with strict inequality if $a \neq \tilde{a}$. Using this with $a = Y - X\hat{\beta}^\lambda$ and $\tilde{a} = Y - X\tilde{\beta}^\lambda$, and invoking the convexity of the norms $\|\cdot\|_{q_j}$, we find

$$\begin{aligned} &g(\|Y - X(\alpha\hat{\beta}^\lambda + (1 - \alpha)\tilde{\beta}^\lambda)\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j(\alpha\hat{\beta}^\lambda + (1 - \alpha)\tilde{\beta}^\lambda)\|_{q_j} \\ &= g(\|\alpha(Y - X\hat{\beta}^\lambda) + (1 - \alpha)(Y - X\tilde{\beta}^\lambda)\|_2^2) + \sum_{j=1}^k \lambda_j \|\alpha M_j\hat{\beta}^\lambda + (1 - \alpha)M_j\tilde{\beta}^\lambda\|_{q_j} \\ &\leq \alpha g(\|Y - X\hat{\beta}^\lambda\|_2^2) + (1 - \alpha)g(\|Y - X\tilde{\beta}^\lambda\|_2^2) + \alpha \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} + (1 - \alpha) \sum_{j=1}^k \lambda_j \|M_j\tilde{\beta}^\lambda\|_{q_j} \\ &= \alpha \left(g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} \right) + (1 - \alpha) \left(g(\|Y - X\tilde{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\tilde{\beta}^\lambda\|_{q_j} \right) \end{aligned}$$

with strict inequality if $X\hat{\beta}^\lambda \neq X\tilde{\beta}^\lambda$. Moreover, we note that

$$\hat{\beta}^\lambda, \tilde{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\beta\|_{q_j} \right\}$$

implies

$$g(\|Y - X\tilde{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\tilde{\beta}^\lambda\|_{q_j} = g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j}.$$

Combining the results yields

$$\begin{aligned} & g(\|Y - X(\alpha\hat{\beta}^\lambda + (1-\alpha)\tilde{\beta}^\lambda)\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j(\alpha\hat{\beta}^\lambda + (1-\alpha)\tilde{\beta}^\lambda)\|_{q_j} \\ & \leq g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} \end{aligned}$$

with strict inequality if $X\hat{\beta}^\lambda \neq X\tilde{\beta}^\lambda$. Using again that

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\beta\|_{q_j} \right\},$$

we find that the above inequality is actually an equality, so that

$$X\hat{\beta}^\lambda = X\tilde{\beta}^\lambda.$$

and

$$\alpha\hat{\beta}^\lambda + (1-\alpha)\tilde{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\beta\|_{q_j} \right\}$$

as desired. □

Proof of Lemma A.2. To show that the function $\lambda \mapsto g(\|Y - X\hat{\beta}^\lambda\|_2^2)$ is continuous, we first show that the function

$$\begin{aligned} & \mathbb{R}^k \rightarrow \mathbb{R} \\ & \lambda \mapsto g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} \end{aligned}$$

is continuous.

It follows from Criterion (2) that

$$g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} \leq g(\|Y\|_2^2).$$

In particular, since the function g is non-negative on $[0, \infty)$ by assumption, it holds that

$$\sum_{j=1}^k \lambda_j \|M_j\hat{\beta}^\lambda\|_{q_j} \leq g(\|Y\|_2^2).$$

Thus, we have

$$\lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} \leq g(\|Y\|_2^2) \quad \text{for } j \in \{1, 2, \dots, k\}.$$

Hence, if $\lambda_j > 0$, it holds that

$$\|M_j \hat{\beta}^\lambda\|_{q_j} \leq \frac{g(\|Y\|_2^2)}{\lambda_j} \quad \text{for } j \in \{1, 2, \dots, k\}. \quad (5)$$

Also note that by Criterion (2), it holds for any pair of tuning parameters $\lambda', \lambda \in (0, \infty)^k$

$$g(\|Y - X \hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} \leq g(\|Y - X \hat{\beta}^{\lambda'}\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^{\lambda'}\|_{q_j}$$

and

$$g(\|Y - X \hat{\beta}^{\lambda'}\|_2^2) + \sum_{j=1}^k \lambda'_j \|M_j \hat{\beta}^{\lambda'}\|_{q_j} \leq g(\|Y - X \hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda'_j \|M_j \hat{\beta}^\lambda\|_{q_j}.$$

Rearranging these two inequalities, we obtain

$$g(\|Y - X \hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} - g(\|Y - X \hat{\beta}^{\lambda'}\|_2^2) - \sum_{j=1}^k \lambda'_j \|M_j \hat{\beta}^{\lambda'}\|_{q_j} \leq \sum_{j=1}^k (\lambda_j - \lambda'_j) \|M_j \hat{\beta}^{\lambda'}\|_{q_j}$$

and

$$g(\|Y - X \hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} - g(\|Y - X \hat{\beta}^{\lambda'}\|_2^2) - \sum_{j=1}^k \lambda'_j \|M_j \hat{\beta}^{\lambda'}\|_{q_j} \geq \sum_{j=1}^k (\lambda_j - \lambda'_j) \|M_j \hat{\beta}^\lambda\|_{q_j}.$$

By Hölder's inequality and Inequality (5), it holds that

$$\begin{aligned} \sum_{j=1}^k (\lambda_j - \lambda'_j) \|M_j \hat{\beta}^\lambda\|_{q_j} &\geq -\|\lambda - \lambda'\|_1 \max_{1 \leq j \leq k} \{ \|M_j \hat{\beta}^\lambda\|_{q_j} \} \\ &\geq -\|\lambda - \lambda'\|_1 \max_{1 \leq j \leq k} \left\{ \frac{g(\|Y\|_2^2)}{\lambda_j} \right\}. \end{aligned}$$

For $\lambda' \in (0, \infty)^k$, we deduce similarly

$$\sum_{j=1}^k (\lambda_j - \lambda'_j) \|M_j \hat{\beta}^{\lambda'}\|_{q_j} \leq \|\lambda - \lambda'\|_1 \max_{1 \leq j \leq k} \left\{ \frac{g(\|Y\|_2^2)}{\lambda'_j} \right\}.$$

Let $m := \max_{1 \leq j \leq k} \left\{ \frac{g(\|Y\|_2^2)}{\lambda'_j} \right\} + \max_{1 \leq j \leq k} \left\{ \frac{g(\|Y\|_2^2)}{\lambda_j} \right\}$. Then, for any $\lambda, \lambda' \in (0, \infty)^k$ such that

$$\|\lambda - \lambda'\|_2 < \frac{\epsilon}{\sqrt{km}}$$

for given $\epsilon > 0$, it holds that

$$\left| g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} - g(\|Y - X\hat{\beta}^{\lambda'}\|_2^2) - \sum_{j=1}^k \lambda'_j \|M_j \hat{\beta}^{\lambda'}\|_{q_j} \right| < \epsilon.$$

This implies the desired continuity.

Now we show that function $\lambda \mapsto g(\|Y - X\hat{\beta}^\lambda\|_2^2)$ is continuous. We proceed with contradiction. Thus, we assume there exist $\lambda' \in (0, \infty)^k$ and $\epsilon_0 > 0$, so that for any $\delta > 0$, there exists $\lambda \in (0, \infty)^k$ satisfying

$$\|\lambda - \lambda'\|_2 < \delta \quad \text{and} \quad \left| g(\|Y - X\hat{\beta}^\lambda\|_2^2) - g(\|Y - X\hat{\beta}^{\lambda'}\|_2^2) \right| \geq \epsilon_0.$$

We note that by Lemma A.1, the value of $g(\|Y - X\hat{\beta}^\lambda\|_2^2)$ does not depend on the specific choice of the estimator $\hat{\beta}^\lambda$. Since $x \mapsto g(x)$ is strictly increasing, there exists $\epsilon'_0 \equiv \epsilon'_0(\epsilon_0) > 0$ such that

$$\left| \|X\hat{\beta}^\lambda\|_2^2 - \|X\hat{\beta}^{\lambda'}\|_2^2 \right| > \epsilon'_0. \quad (6)$$

Define the set $B := \left\{ \beta \in \mathbb{R}^p : \|X\hat{\beta}^{\lambda'}\|_2^2 - \|X\beta\|_2^2 > \epsilon'_0 \right\}$. It follows directly that $\hat{\beta}^{\lambda'} \notin B$, and due to Inequality (6) above, it follows that $\hat{\beta}^\lambda \in B$. Let $\eta > 0$ be arbitrary. Invoking Criterion (2) and the continuity shown above, we obtain

$$\begin{aligned} & g(\|Y - X\hat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \hat{\beta}^\lambda\|_{q_j} \\ &= \min_{\beta \in B} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\} \\ &\geq \min_{\beta \in B} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda'_j \|M_j \beta\|_{q_j} \right\} - \eta \end{aligned}$$

if δ is sufficiently small. Moreover, using $\hat{\beta}^{\lambda'} \notin B$ and again the continuity, it holds for δ sufficiently small that

$$\begin{aligned} & \min_{\beta \in B} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda'_j \|M_j \beta\|_{q_j} \right\} \\ &> \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda'_j \|M_j \beta\|_{q_j} \right\} + \xi \\ &\geq \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\} + \xi/2 \end{aligned}$$

for a $\xi \equiv \xi(\lambda', \epsilon'_0) > 0$. Choosing $\eta = \xi/4$, we find

$$\begin{aligned} & \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\} \\ & > \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\} + \xi/4, \end{aligned}$$

which is a contradiction and thus concludes the proof of the continuity of the function $\lambda \mapsto g(\|Y - X\hat{\beta}^\lambda\|_2^2)$. The continuity of the function $\lambda \mapsto \|Y - X\hat{\beta}^\lambda\|_2^2$ then follows from the assumption that the link function g is continuous and increasing. This concludes the proof of the lemma. \square

Proof of Lemma A.3. Since $Y \neq \mathbf{0}_n$ with probability one, we assume $Y \neq \mathbf{0}_n$ in the following.

We then show that

$$\|Y - X\hat{\beta}^\lambda\|_2 > 0.$$

We do this by contradiction, that is, we assume

$$\|Y - X\hat{\beta}^\lambda\|_2 = 0.$$

This implies

$$Y - X\hat{\beta}^\lambda = \mathbf{0}_n. \quad (7)$$

Since $\beta \mapsto g(\|Y - X\beta\|_2^2)$ is convex, the subdifferential $\partial_\beta g(\|Y - X\beta\|_2^2) \Big|_{\beta=\hat{\beta}^\lambda}$ exists. Thus, the KKT conditions imply

$$\mathbf{0}_p \in \partial_\beta g(\|Y - X\beta\|_2^2) \Big|_{\beta=\hat{\beta}^\lambda} + \sum_{j=1}^k \lambda_j \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\hat{\beta}^\lambda},$$

which implies by the chain rule

$$\mathbf{0}_p \in \partial_x g(x) \Big|_{x=\|Y - X\hat{\beta}^\lambda\|_2^2} (-2X^\top (Y - X\hat{\beta}^\lambda)) + \sum_{j=1}^k \lambda_j \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\hat{\beta}^\lambda}.$$

Plugging Equality (7) into this display yields

$$\mathbf{0}_p \in \sum_{j=1}^k \lambda_j \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\hat{\beta}^\lambda}.$$

This means that the vector $\hat{\beta}^\lambda$ minimizes the function $\beta \mapsto \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j}$. However, since $\lambda_1, \dots, \lambda_k > 0$, and by assumption on the matrices M_1, \dots, M_k , this mapping is a

norm and is thus minimized only at $\mathbf{0}_p$. Consequently, $\hat{\beta}^\lambda = \mathbf{0}_p$. However, Equality (7) then gives

$$Y = X\hat{\beta}^\lambda = X\mathbf{0}_p = \mathbf{0}_n,$$

which contradicts $Y \neq \mathbf{0}_n$. Thus, $Y - X\hat{\beta}^\lambda \neq \mathbf{0}_n$, and it follows that $\|Y - X\hat{\beta}^\lambda\|_2^2 \neq 0$.

Since the function $x \mapsto g(x)$ is continuously differentiable on $(0, \infty)$ with strictly positive derivative, we finally obtain

$$g'(\|Y - X\hat{\beta}^\lambda\|_2^2) > 0.$$

as desired. \square

Proof of Lemma A.4. We first note that $\mathbf{0}_p \in \partial_\beta f(\beta) \Big|_{\beta=\bar{\beta}}$ implies that $\bar{\beta}$ is a minimizer of the function f . Consider now a vector $\tilde{\beta} \in \mathbb{R}^p$ and the function

$$\begin{aligned} [0, 1] &\rightarrow \mathbb{R} \\ \alpha &\mapsto h(\alpha) := \alpha\tilde{\beta} + (1 - \alpha)\bar{\beta}. \end{aligned}$$

If $\tilde{\beta} = \bar{\beta}$, it trivially holds that $\kappa^\top(\tilde{\beta} - \bar{\beta}) \geq 0$. More generally, assume there is a $c < 0$ such that

$$c \in \frac{\partial}{\partial \alpha} f(h(\alpha)) \Big|_{\alpha=0} = \left(\partial_\beta f(\beta) \Big|_{\beta=\bar{\beta}} \right)^\top (\tilde{\beta} - \bar{\beta}).$$

However, since every convex function is continuous, then for a small value of α , it holds that $f(h(\alpha)) < f(\bar{\beta})$. This contradicts that $\bar{\beta}$ is a minimizer of the function f . Thus, for all $\kappa \in \partial_\beta f(\beta) \Big|_{\beta=\bar{\beta}}$ and $\tilde{\beta} \in \mathbb{R}^p$, it holds that

$$\kappa^\top(\tilde{\beta} - \bar{\beta}) \geq 0$$

as desired. \square

A.3. Proof of Lemma 2.1

Proof of Lemma 2.1. The proof consists of three steps. First, we show that the solution equals zero if the tuning parameters are large enough. Second, we show that if one element of the tuning parameter is sufficiently large, increasing that element does not affect the estimator. Finally, we use these results to show the existence of an oracle tuning parameter.

Let us start with some notation. For each $j \in \{1, \dots, k\}$, we define the set $A_j \subset \{1, 2, \dots, p\}$ such that for any $u \in A_j$, the u th row of M_j is not zero, that is, $A_j := \{u \in \{1, \dots, p\} : \max_{v \in \{1, \dots, p\}} |(M_j)_{uv}| > 0\}$. By assumption on the sequence M_1, \dots, M_k , it holds that $\bigcup_{j=1}^k A_j = \{1, \dots, p\}$. For $j \in \{1, \dots, k\}$, define the vector $(X^\top Y)^j \in \mathbb{R}^p$ via $(X^\top Y)_i^j := (X^\top Y)_i \mathbb{1}\{i \in A_j\}$, and set $m := \max_{j \in \{1, \dots, k\}} 2g'(\|Y\|_2^2) \|(X^\top Y)^j\|_{q_j}^* \vee 1$, where \vee denotes the maximum.

The three mentioned steps now read in detail:

1. Show that for any $\lambda \in [m, \infty)^k$, it holds that

$$\{\mathbf{0}_p\} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

2. Show that $X\widehat{\beta}^{\tilde{\lambda}} = X\widehat{\beta}^{\lambda}$ if $\tilde{\lambda}, \lambda \in (0, \infty)^k$ satisfy

$$\begin{aligned} \tilde{\lambda}_j &> \lambda_j = m && \text{if } j \in B \\ \tilde{\lambda}_j &= \lambda_j && \text{if } j \notin B \end{aligned}$$

for a non-empty subset $B \subset \{1, 2, \dots, k\}$.

3. Show that with probability one, there exists a vector $\lambda \in (0, \infty)^k$ that satisfies

$$\frac{\lambda_j}{2g'(\|Y - X\widehat{\beta}^{\lambda}\|_2^2)} = \|(XP_j M_j^+)^{\top} \varepsilon\|_{q_j}^* \quad (8)$$

for all $j \in \{1, \dots, k\}$.

Step 1: We first show that

$$\mathbf{0}_p \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}$$

implies

$$\{\mathbf{0}_p\} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

Assume for a $\lambda \in [m, \infty)^k$, it holds that

$$\mathbf{0}_p \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

then by Lemma A.1, for any $\widehat{\beta}^{\lambda} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}$, it holds that $X\widehat{\beta}^{\lambda} = \mathbf{0}_n$. Hence, by assumption on the matrices M_1, \dots, M_k , it holds for

any $\widehat{\beta}^\lambda \neq \mathbf{0}_p$,

$$\begin{aligned}
& g(\|Y - X\widehat{\beta}^\lambda\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \widehat{\beta}^\lambda\|_{q_j} \\
&= g(\|Y\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \widehat{\beta}^\lambda\|_{q_j} \\
&> g(\|Y\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \mathbf{0}_p\|_{q_j} \\
&= g(\|Y - X\mathbf{0}_p\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \mathbf{0}_p\|_{q_j}.
\end{aligned}$$

This contradicts $\widehat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \{g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j}\}$, and thus, $\widehat{\beta}^\lambda = \mathbf{0}_p$. Thus,

$$\{\mathbf{0}_p\} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}$$

as desired. It is left to show that for any vector $\lambda \in [m, \infty)^k$, it holds that

$$\mathbf{0}_p \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

By the KKT conditions, we have to show that there are vectors $\kappa(j) \in \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\mathbf{0}_p}$ such that

$$-2g'(\|Y\|_2^2)X^\top Y + \sum_{j=1}^k \lambda_j \kappa(j) = \mathbf{0}_p. \quad (9)$$

Define $\tilde{A}_1 := A_1$, $\tilde{A}_j := A_j \setminus \{A_1, \dots, A_{j-1}\}$ for $j = 2, \dots, k$. In particular, $\tilde{A}_j \subset A_j$, the \tilde{A}_j 's are disjoint, and $\bigcup_{j=1}^k \tilde{A}_j = \{1, \dots, p\}$. With this notation, we need to show that for any $v \in \tilde{A}_j$ and $j \in \{1, \dots, k\}$, it holds that

$$-2g'(\|Y\|_2^2)(X^\top Y)_v + \sum_{j=1}^k \lambda_j \kappa(j)_v = 0.$$

Define the vector $\kappa(j)$ for $j \in \{1, \dots, k\}$ via

$$\kappa(j)_v := \begin{cases} \frac{2g'(\|Y\|_2^2)(X^\top Y)_v}{\lambda_j} & \text{if } v \in \tilde{A}_j \\ 0 & \text{if } v \notin \tilde{A}_j. \end{cases}$$

Since $\tilde{A}_j \subset A_j$ and \tilde{A}_j 's are disjoint, we find for all $j \in \{1, \dots, k\}$ and $v \in \tilde{A}_j$, that

$$-2g'(\|Y\|_2^2)(X^\top Y)_v + \sum_{j=1}^k \lambda_j \kappa(j)_v = -2g'(\|Y\|_2^2)(X^\top Y)_v^j + \lambda_j \kappa(j)_v.$$

By definition of the vectors $\kappa(j)$ for $j \in \{1, \dots, k\}$, we thus have

$$-2g'(\|Y\|_2^2)(X^\top Y)_v^j + \lambda_j \kappa(j)_v = 0,$$

which implies

$$-2g'(\|Y\|_2^2)(X^\top Y)_v + \sum_{j=1}^k \lambda_j \kappa(j)_v = 0.$$

Since $\lambda_j \geq m \geq 2g'(\|Y\|_2^2)\|(X^\top Y)^j\|_{q_j}^*$, it also follows by taking the dual norm on both sides that

$$\|\kappa(j)\|_{q_j}^* \leq \frac{2g'(\|Y\|_2^2)\|(X^\top Y)^j\|_{q_j}^*}{\lambda_j} \leq 1, \quad (10)$$

and hence, $\kappa(j) \in \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\mathbf{0}_p}$ for all $j \in \{1, \dots, k\}$. So we get

$$\mathbf{0}_p \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

as desired. We conclude that

$$\{\mathbf{0}_p\} = \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}.$$

Step 2: Consider a pair of vectors $\lambda, \tilde{\lambda} \in (0, \infty)^k$, that satisfy $\tilde{\lambda}_j > \lambda_j = m$ for $j \in B$ and $\tilde{\lambda}_j = \lambda_j$ for $j \in \{1, \dots, k\} \setminus B$. For λ , fix a solution

$$\hat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j} \right\}$$

with corresponding subdifferentials $\kappa(j) \in \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\hat{\beta}^\lambda}$ that satisfy the KKT conditions

$$-2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)X^\top(Y - X\hat{\beta}^\lambda) + \sum_{j=1}^k \lambda_j \kappa(j) = \mathbf{0}_p.$$

We first need to show that

$$\widehat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2) + \sum_{j=1}^k \tilde{\lambda}_j \|M_j \beta\|_{q_j} \right\}.$$

By the KKT conditions, we have to show that there are the vectors $\tilde{\kappa}(j) \in \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\widehat{\beta}^\lambda}$ such that

$$-2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)X^\top(Y - X\widehat{\beta}^\lambda) + \sum_{j=1}^k \tilde{\lambda}_j \tilde{\kappa}(j) = \mathbf{0}_p.$$

Define $\tilde{\kappa}(j)$ for $j \in \{1, \dots, k\}$ via

$$\tilde{\kappa}(j) := \frac{\lambda_j}{\tilde{\lambda}_j} \kappa(j).$$

Plugging this into the previous display yields

$$-2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)X^\top(Y - X\widehat{\beta}^\lambda) + \sum_{j=1}^k \tilde{\lambda}_j \tilde{\kappa}(j) = -2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)X^\top(Y - X\widehat{\beta}^\lambda) + \sum_{j=1}^k \lambda_j \kappa_j.$$

Therefore, it holds that

$$-2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)X^\top(Y - X\widehat{\beta}^\lambda) + \sum_{j=1}^k \tilde{\lambda}_j \tilde{\kappa}(j) = \mathbf{0}_p.$$

Moreover, by definition of $\tilde{\kappa}(j)$ and Inequality (10), we have that

$$\|\tilde{\kappa}(j)\|_{q_j}^* = \frac{\lambda_j}{\tilde{\lambda}_j} \|\kappa(j)\|_{q_j}^* \leq \|\kappa(j)\|_{q_j}^* \leq 1,$$

for all $j \in B$. So, $\tilde{\kappa}(j) \in \partial_\beta \|M_j \beta\|_{q_j} \Big|_{\beta=\widehat{\beta}^\lambda}$ for any $j \in \{1, \dots, k\}$. Hence, it holds that

$$\widehat{\beta}^\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ g(\|Y - X\beta\|_2) + \sum_{j=1}^k \tilde{\lambda}_j \|(M_j \beta)^j\|_{q_j} \right\}.$$

This gives $X\widehat{\beta}^\lambda = X\widehat{\beta}^\lambda$ by Lemma A.1.

Step 3: Finally, we show the existence of a $\lambda \in (0, \infty)^k$ that satisfies Equality (8). For this, we define

$$a := m \wedge \left(2g'(\|Y\|_2^2) \min_{j \in \{1, \dots, k\}} \|(XP_j M_j^+)^\top \varepsilon\|_{q_j}^* \right),$$

where \wedge denotes the minimum. By assumption on the noise ε , it holds that $a > 0$ with probability one. Next, we consider the function

$$f : [a, m]^k \rightarrow \mathbb{R}^k$$

$$\lambda \mapsto f(\lambda) := 2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(\|(XP_1 M_1^+)^\top \varepsilon\|_{q_1}^*, \dots, \|(XP_k M_k^+)^\top \varepsilon\|_{q_k}^* \right)^\top.$$

Note first that $g(\|Y - X\hat{\beta}^\lambda\|_2^2) \leq g(\|Y\|_2^2)$ by definition of the estimator $\hat{\beta}^\lambda$. Hence, since g is increasing and g' is non-increasing, we find

$$\min_{\lambda \in [a, m]^k} \min_{j \in \{1, \dots, k\}} f(\lambda)_j \geq a.$$

Note also that since $[a, m]^k$ is compact, it holds that

$$\sup_{\lambda \in [a, m]^k} \|f(\lambda)\|_\infty \leq b$$

for some $b \in (0, \infty)$. Using this and Step 2, we find that the function f and the function

$$h : [a, b \vee m]^k \rightarrow [a, b \vee m]^k$$

$$\lambda \mapsto h(\lambda) := 2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(\|(XP_1 M_1^+)^\top \varepsilon\|_{q_1}^*, \dots, \|(XP_k M_k^+)^\top \varepsilon\|_{q_k}^* \right)^\top$$

have the same images, that is,

$$\{y : y = f(\lambda), \lambda \in [a, m]^k\} = \{y : y = h(\lambda), \lambda \in [a, b \vee m]^k\}.$$

By Lemma A.2 and the assumption that $g'(x)$ is continuous on $[0, \infty)$, the function h is continuous. Moreover, $[a, b \vee m]^k$ is a compact and convex subset of \mathbb{R}^k . We can thus apply Brouwer's fixed-point theorem to deduce that

$$\lambda = 2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(\|(XP_1 M_1^+)^\top \varepsilon\|_{q_1}^*, \dots, \|(XP_k M_k^+)^\top \varepsilon\|_{q_k}^* \right)^\top$$

for a vector $\lambda \in [a, b \vee m]^k$. According to Lemma A.3, it holds that $g'(\|Y - X\hat{\beta}^\lambda\|_2^2) > 0$ with probability one, so that

$$\frac{\lambda}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} = \left(\|(XP_1 M_1^+)^\top \varepsilon\|_{q_1}^*, \dots, \|(XP_k M_k^+)^\top \varepsilon\|_{q_k}^* \right)^\top$$

as desired. \square

A.4. Proof of Theorem 2.1

Proof of Theorem 2.1. Consider the function

$$\mathbb{R}^p \rightarrow \mathbb{R}$$

$$\beta \mapsto f(\beta) := g(\|Y - X\beta\|_2^2) + \sum_{j=1}^k \lambda_j \|M_j \beta\|_{q_j}.$$

According to the assumption on the link function g , the function $\beta \mapsto f(\beta)$ is convex. Let $\hat{\beta}^\lambda$ be a minimizer of the function f for a fixed $\lambda \in (0, \infty)^k$. By Lemma A.4, for every $\kappa \in \partial f_\beta(\beta)|_{\beta=\hat{\beta}^\lambda}$ and any $\tilde{\beta} \in \mathbb{R}^p$, it holds that

$$\kappa^\top (\tilde{\beta} - \hat{\beta}^\lambda) \geq 0.$$

Let us calculate the left-hand side of this inequality. By Lemma A.3 and the assumption that the function g is differentiable on $(0, \infty)$, it holds that

$$\left. \frac{\partial}{\partial \beta} g(\|Y - X\beta\|_2^2) \right|_{\beta=\hat{\beta}^\lambda} = -2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(X^\top (Y - X\hat{\beta}^\lambda) \right).$$

Moreover, by definition of subdifferentials, it holds that $\kappa_j \in \partial\{\alpha \mapsto \|M_j^j \alpha\|_{q_j}\}|_{\alpha=\hat{\beta}^\lambda}$ if and only if for all $\tilde{\beta} \in \mathbb{R}^p$

$$\|M_j \tilde{\beta}\|_{q_j} \geq \|M_j \hat{\beta}^\lambda\|_{q_j} + \langle \kappa_j, \tilde{\beta} - \hat{\beta}^\lambda \rangle,$$

which is equivalent to

$$\|M_j \tilde{\beta}\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j} \geq \kappa_j^\top (\tilde{\beta} - \hat{\beta}^\lambda).$$

Combining these results yields

$$-2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) (X^\top (Y - X\hat{\beta}^\lambda))^\top (\tilde{\beta} - \hat{\beta}^\lambda) + \sum_{j=1}^k \lambda_j (\|M_j \tilde{\beta}\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j}) \geq 0.$$

According to the model (1), we can replace Y with $X\beta^* + \varepsilon$ to obtain

$$\begin{aligned} & (X^\top (Y - X\hat{\beta}^\lambda))^\top (\tilde{\beta} - \hat{\beta}^\lambda) \\ &= (X^\top (X\beta^* + \varepsilon - X\hat{\beta}^\lambda))^\top (\tilde{\beta} - \hat{\beta}^\lambda) \\ &= (X(\beta^* - \hat{\beta}^\lambda))^\top X(\tilde{\beta} - \hat{\beta}^\lambda) + \varepsilon^\top X(\tilde{\beta} - \hat{\beta}^\lambda) \\ &= \|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 + \langle \varepsilon, X(\beta^* - \hat{\beta}^\lambda) \rangle. \end{aligned}$$

Taking $\tilde{\beta} = \beta^*$ and plugging this back into the previous display yields

$$-2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 + \langle \varepsilon, X(\beta^* - \hat{\beta}^\lambda) \rangle \right) + \sum_{j=1}^k \lambda_j (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j}) \geq 0.$$

Rearranging this inequality, we obtain

$$2g'(\|Y - X\hat{\beta}^\lambda\|_2^2) \left(\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 + \langle \varepsilon, X(\beta^* - \hat{\beta}^\lambda) \rangle \right) \leq \sum_{j=1}^k \lambda_j (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j}).$$

According to Lemma A.3, we can divide both sides by $g'(\|Y - X\hat{\beta}^\lambda\|_2^2)$, so that

$$\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 + \langle \varepsilon, X(\beta^* - \hat{\beta}^\lambda) \rangle \leq \sum_{j=1}^k \frac{\lambda_j}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j})$$

and

$$\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 \leq \langle \varepsilon, X(\hat{\beta}^\lambda - \beta^*) \rangle + \sum_{j=1}^k \frac{\lambda_j}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j})$$

with probability one. Recall that by Equation (3), the vector $\hat{\beta}^\lambda - \beta^*$ can be rewritten as

$$\hat{\beta}^\lambda - \beta^* = \sum_{j=1}^k P_j M_j^+ M_j (\hat{\beta}^\lambda - \beta^*).$$

So we can reorganize the inner product $\langle \varepsilon, X(\hat{\beta}^\lambda - \beta^*) \rangle$ according to

$$\begin{aligned} \langle \varepsilon, X(\hat{\beta}^\lambda - \beta^*) \rangle &= \left\langle \varepsilon, X \sum_{j=1}^k P_j M_j^+ M_j (\hat{\beta}^\lambda - \beta^*) \right\rangle \\ &= \sum_{j=1}^k \left\langle \varepsilon, X P_j M_j^+ M_j (\hat{\beta}^\lambda - \beta^*) \right\rangle \\ &= \sum_{j=1}^k \left\langle (X P_j M_j^+)^T \varepsilon, M_j (\hat{\beta}^\lambda - \beta^*) \right\rangle. \end{aligned}$$

Plugging this into the previous inequality yields

$$\begin{aligned} &\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 \\ &\leq \sum_{j=1}^k \left(\left\langle (X P_j M_j^+)^T \varepsilon, M_j (\hat{\beta}^\lambda - \beta^*) \right\rangle + \frac{\lambda_j}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j}) \right). \end{aligned}$$

Using Hölder's Inequality, we can rewrite the first term according to

$$\left\langle (X P_j M_j^+)^T \varepsilon, M_j (\hat{\beta}^\lambda - \beta^*) \right\rangle \leq \|(X P_j M_j^+)^T \varepsilon\|_{q_j^*} \|M_j \hat{\beta}^\lambda - M_j \beta^*\|_{q_j}.$$

Plugging this back into the previous display gives

$$\begin{aligned} &\|X(\beta^* - \hat{\beta}^\lambda)\|_2^2 \\ &\leq \sum_{j=1}^k \left(\|(X P_j M_j^+)^T \varepsilon\|_{q_j^*} \|M_j \hat{\beta}^\lambda - M_j \beta^*\|_{q_j} + \frac{\lambda_j}{2g'(\|Y - X\hat{\beta}^\lambda\|_2^2)} (\|M_j \beta^*\|_{q_j} - \|M_j \hat{\beta}^\lambda\|_{q_j}) \right). \end{aligned}$$

We can modify this further by applying the triangle inequality and by reorganizing the terms. We find

$$\begin{aligned}
& \|X(\beta^* - \widehat{\beta}^\lambda)\|_2^2 \\
& \leq \sum_{j=1}^k \left(\|(XP_j M_j^+)^T \varepsilon\|_{q_j}^* (\|M_j \widehat{\beta}^\lambda\|_{q_j} + \|M_j \beta^*\|_{q_j}) + \frac{\lambda_j}{2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)} (\|M_j \beta^*\|_{q_j} - \|M_j \widehat{\beta}^\lambda\|_{q_j}) \right) \\
& = \sum_{j=1}^k \left(\|(XP_j M_j^+)^T \varepsilon\|_{q_j}^* + \frac{\lambda_j}{2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)} \right) \|M_j \beta^*\|_{q_j} \\
& \quad + \sum_{j=1}^k \left(\|(XP_j M_j^+)^T \varepsilon\|_{q_j}^* - \frac{\lambda_j}{2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)} \right) \|M_j \widehat{\beta}^\lambda\|_{q_j}.
\end{aligned}$$

We now set $\lambda = \overline{\lambda}$. Recall that by definition of the oracle tuning parameter, it holds that

$$\frac{\overline{\lambda}_j}{2g'(\|Y - X\widehat{\beta}^\lambda\|_2^2)} = \|(XP_j M_j^+)^T \varepsilon\|_{q_j}^*$$

for all $j \in \{1, \dots, k\}$. Using this, the above inequality simplifies to

$$\|X(\beta^* - \overline{\beta})\|_2^2 \leq 2 \sum_{j=1}^k \|(XP_j M_j^+)^T \varepsilon\|_{q_j}^* \|M_j \beta^*\|_{q_j}.$$

To bring this into the standard form, we finally divide both sides by n and find

$$\frac{1}{n} \|X(\beta^* - \overline{\beta})\|_2^2 \leq \frac{2}{n} \sum_{j=1}^k \|(XP_j M_j^+)^T \varepsilon\|_{q_j}^* \|M_j \beta^*\|_{q_j}$$

as desired. □